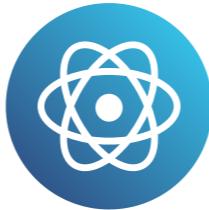


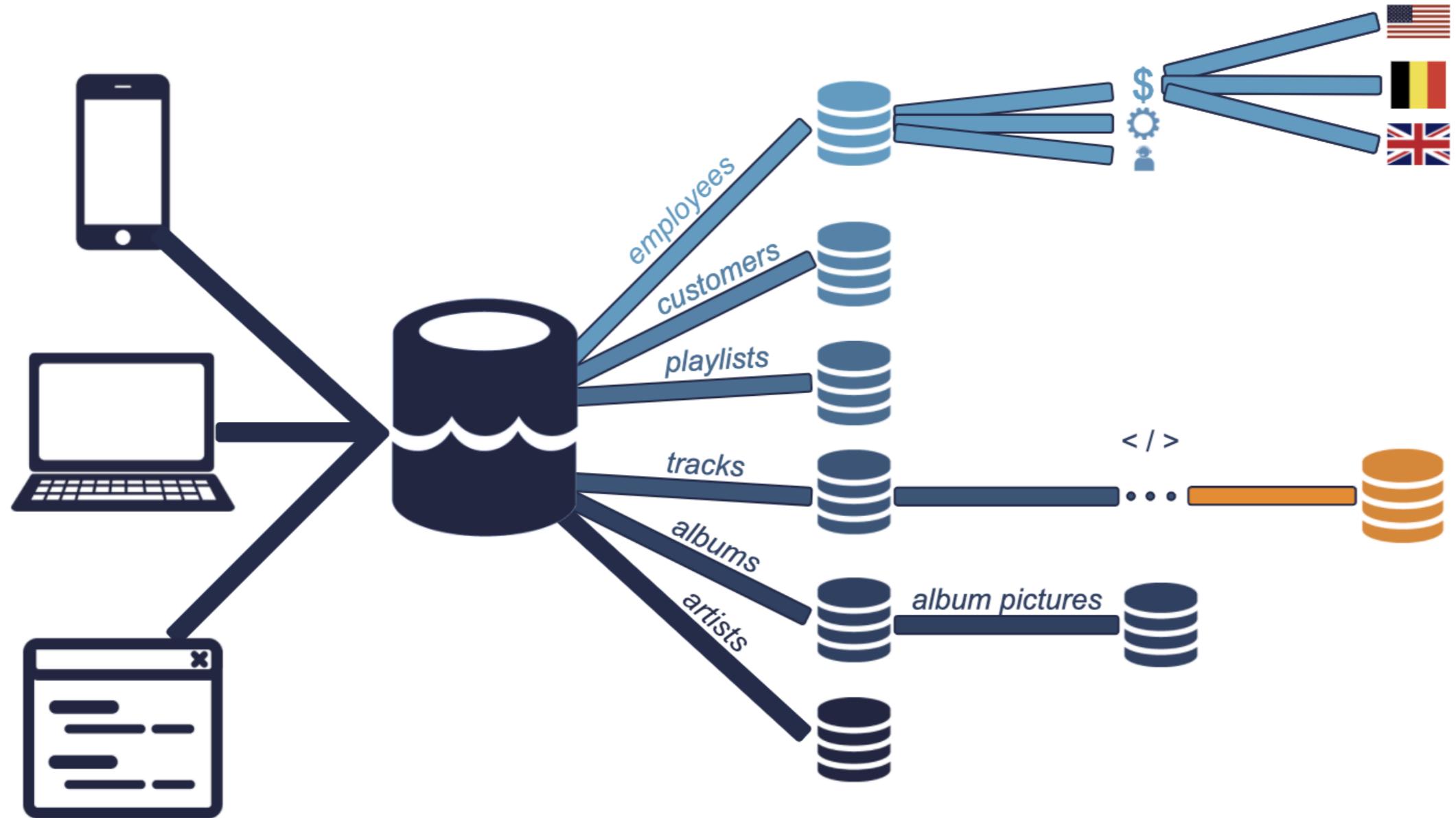
# Processing data

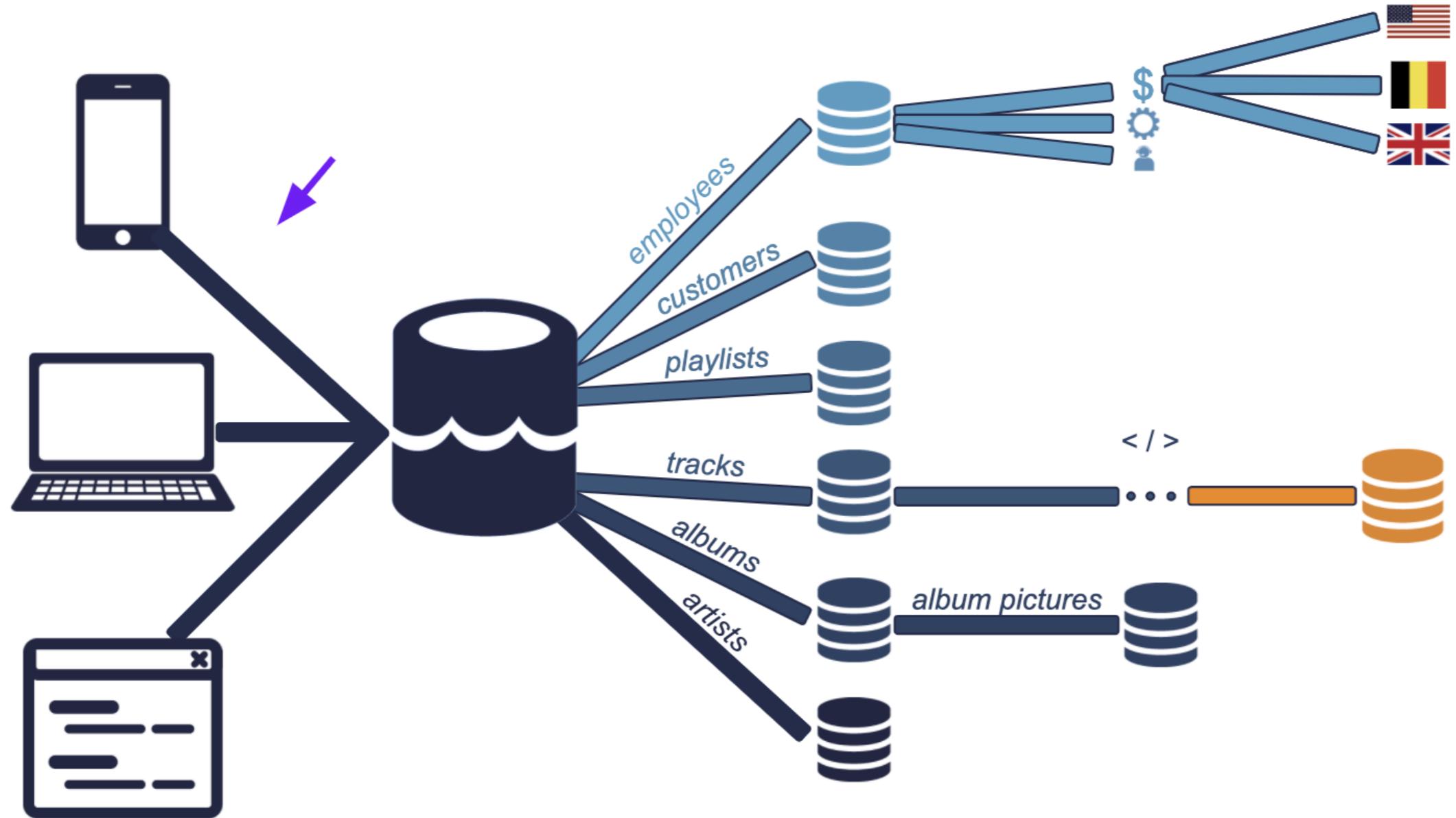
DATA ENGINEERING FOR EVERYONE

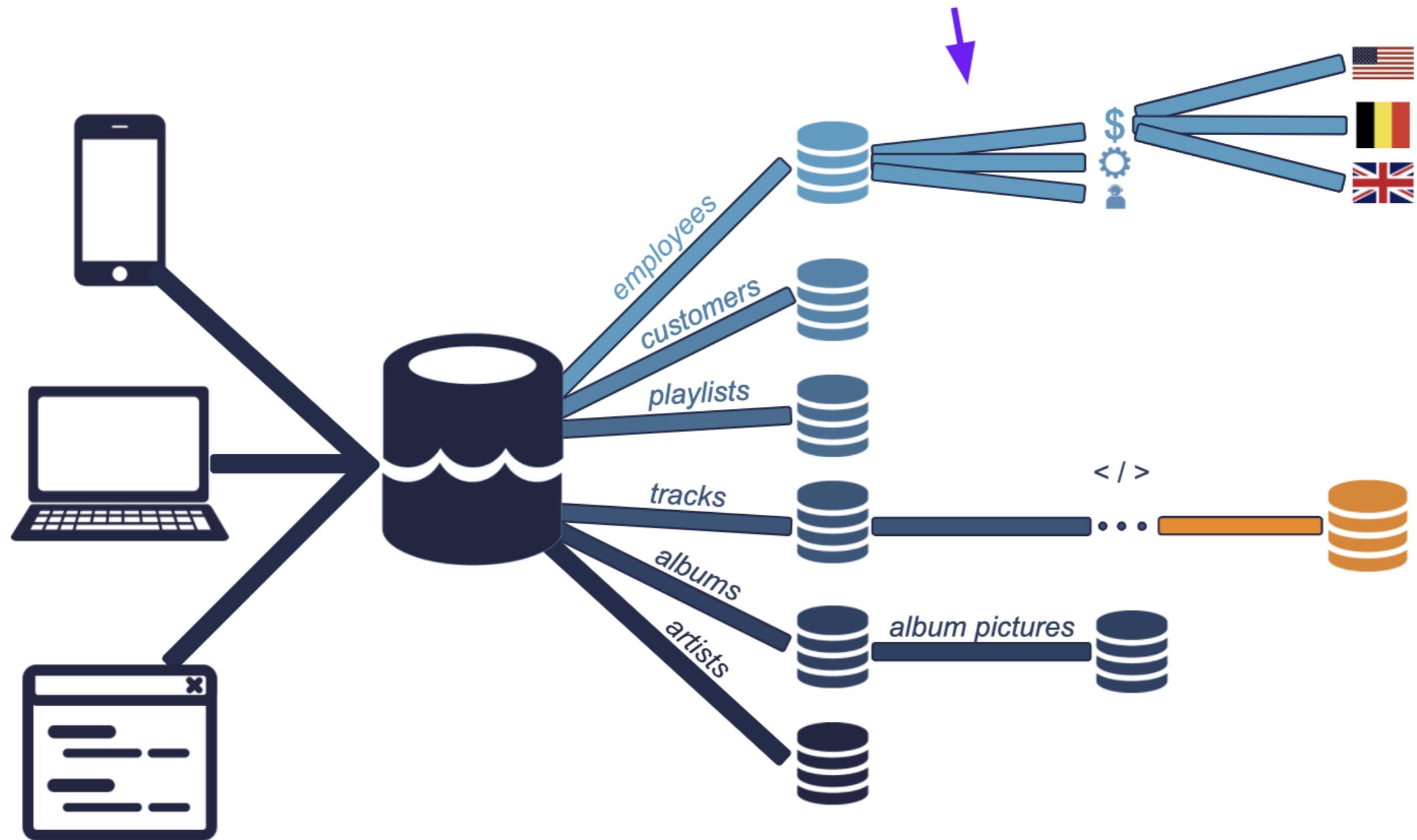


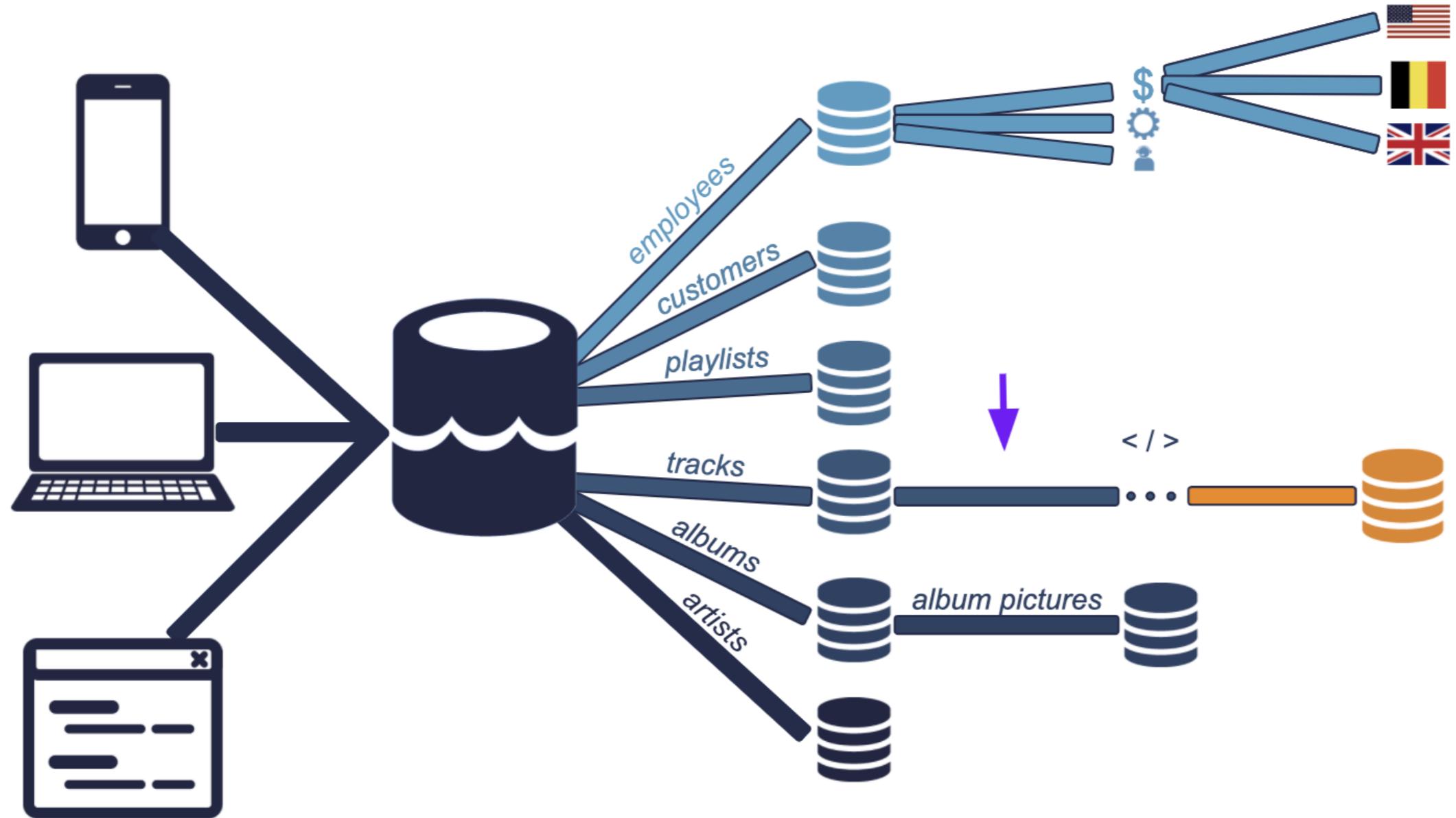
**Hadrien Lacroix**

Content Developer at DataCamp









# A general definition

- Data processing: converting **raw** data into **meaningful** information

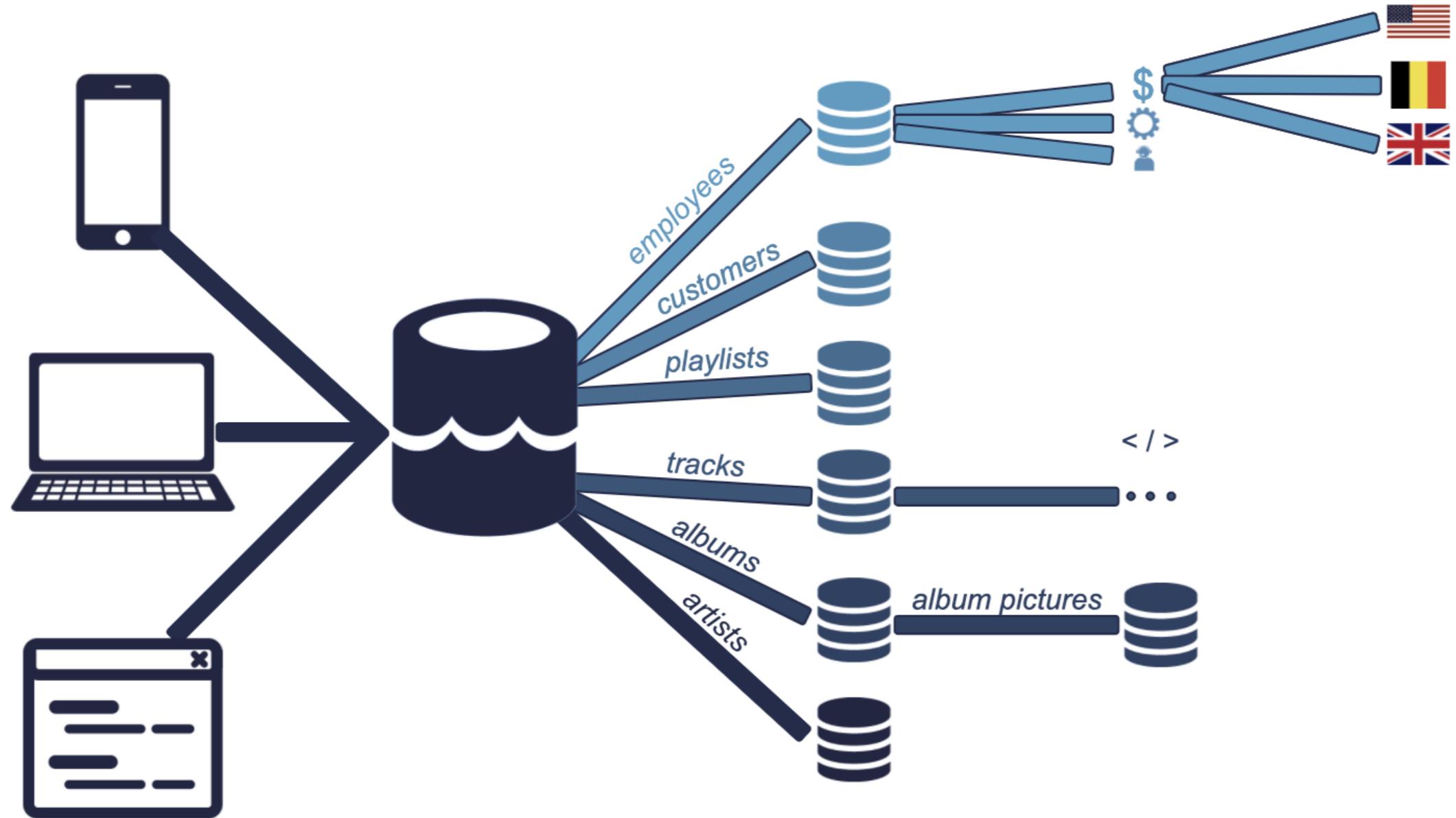
# Data processing value

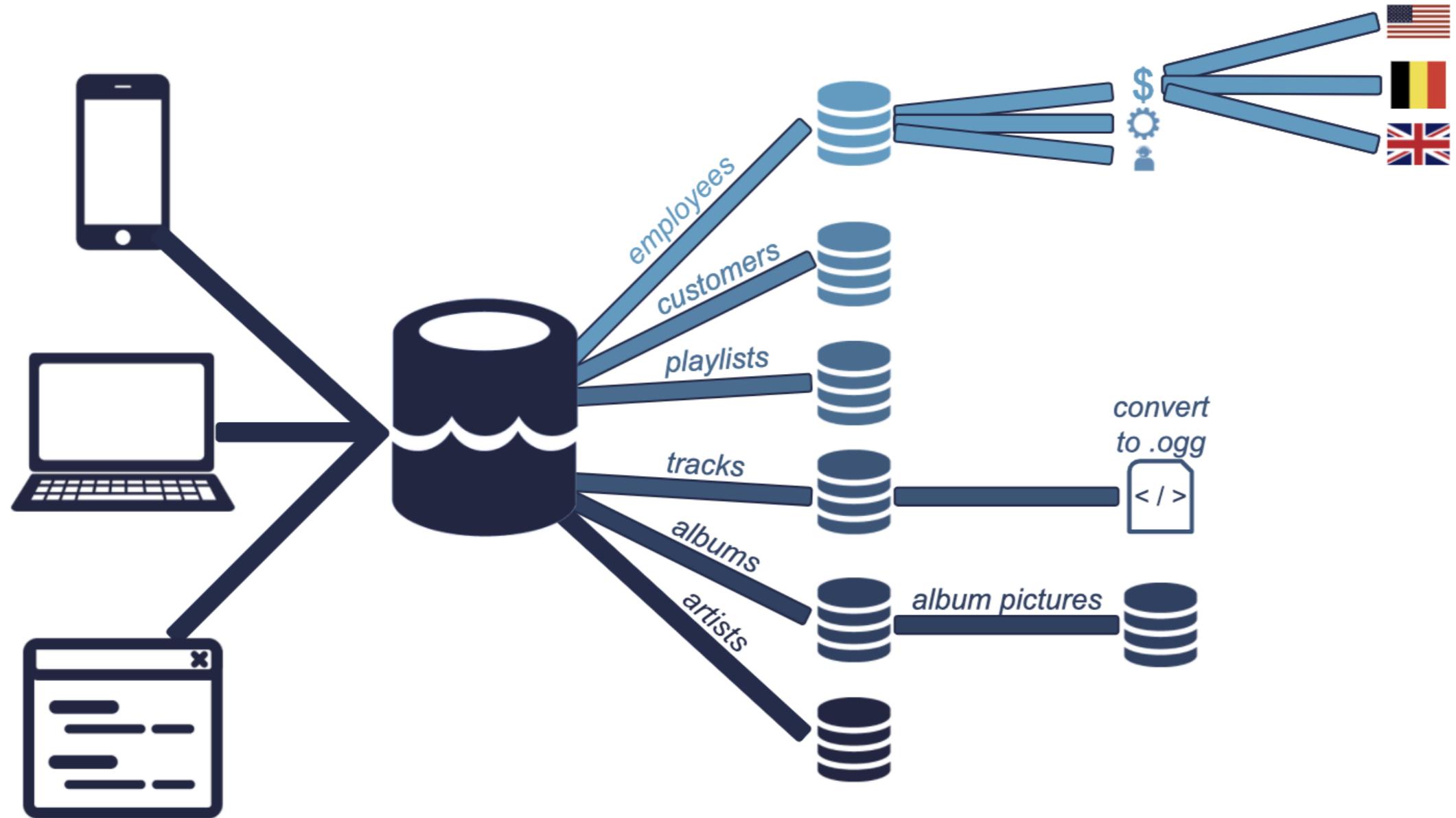
## Conceptually

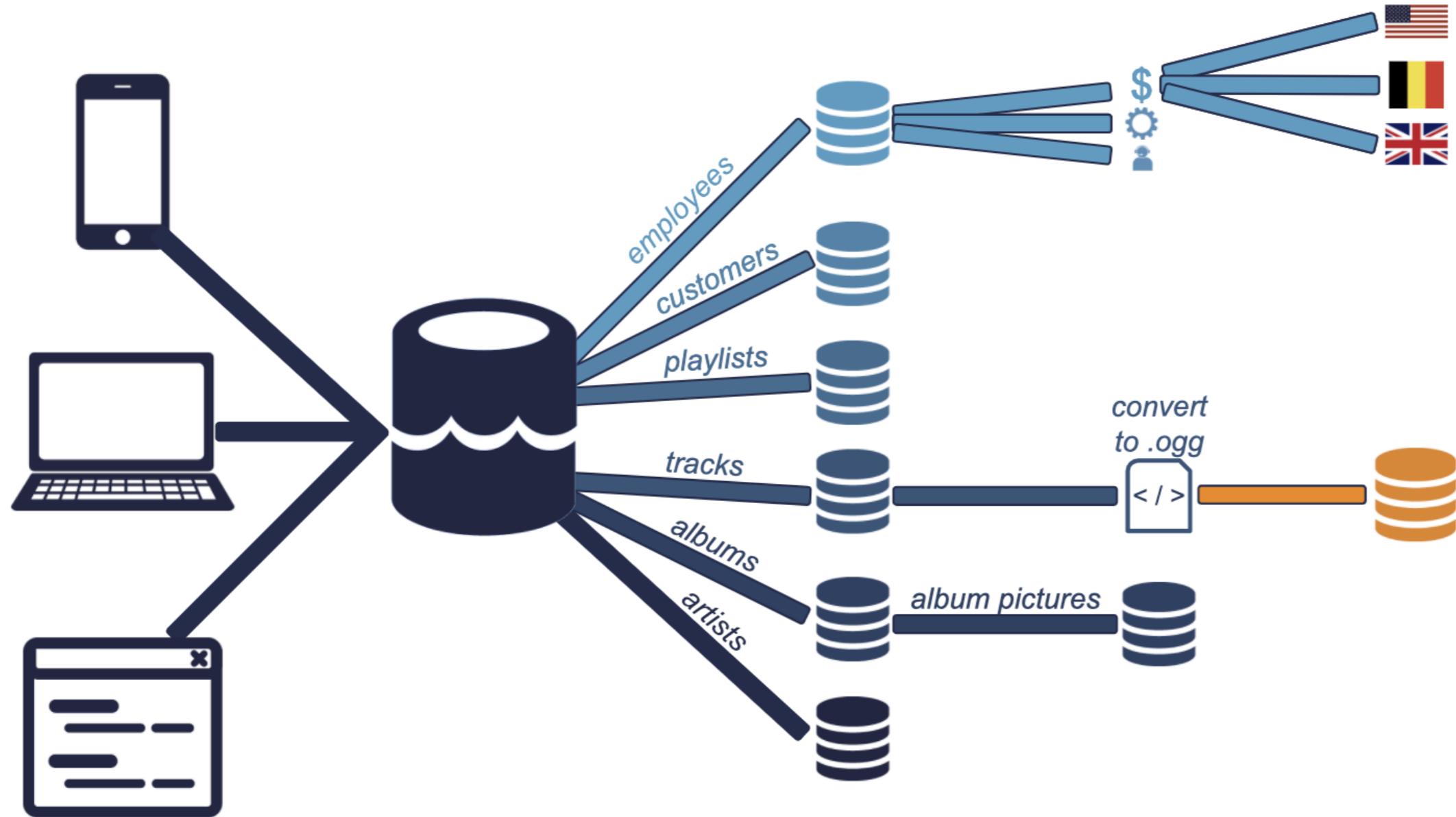
- Remove unwanted data
- Optimize memory, process and network costs
- Convert data from one type to another

## At Spotflix

- No long term need for testing feature data
- Can't afford to store and stream files this big







# Data processing value

## Conceptually

- Remove unwanted data
- To save memory
- Convert data from one type to another
- Organize data
- To fit into a schema/structure
- Increase productivity

## At Spotflix

- No need for lossless format
- Can't afford to store files this big
- Convert songs from `.flac` to `.ogg`
- Reorganize data from the data lake to data warehouses
- Employee table example
- Enable data scientists

# How data engineers process data

- Data manipulation, cleaning, and tidying tasks
  - that can be automated
  - that will always need to be done
- Store data in a sanely structured database
- Create views on top of the database tables
- Optimizing the performance of the database
- Rejecting corrupt song files
- Deciding what happens with missing metadata
- Separate artists and albums tables...
- ...but provide view combining them
- Indexing

# Batch processing



amazon  
EMR



presto



# Stream processing



samza



APACHE  
**STORM™**  
Distributed • Resilient • Real-time



Spring Cloud  
Data Flow



<sup>1</sup> The difference between batch and stream will be explained in the next lesson!



# Summary

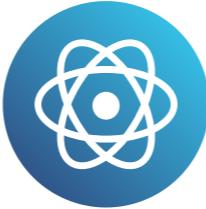
- What data processing is
- Why it's necessary
- What it consists in
- How we process data at Spotflix

# Let's practice!

DATA ENGINEERING FOR EVERYONE

# Scheduling data

DATA ENGINEERING FOR EVERYONE



**Hadrien Lacroix**

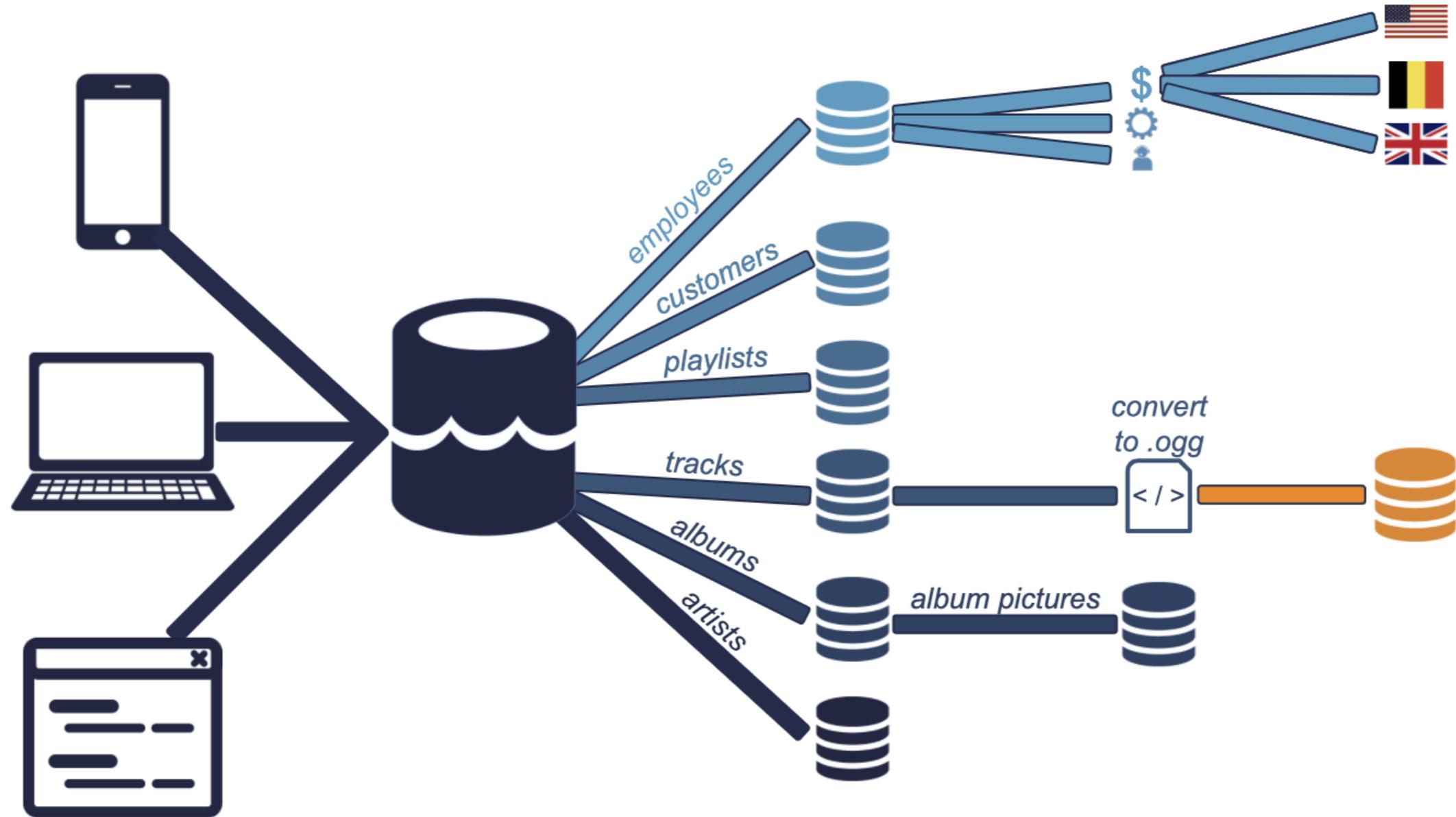
Content Developer at DataCamp

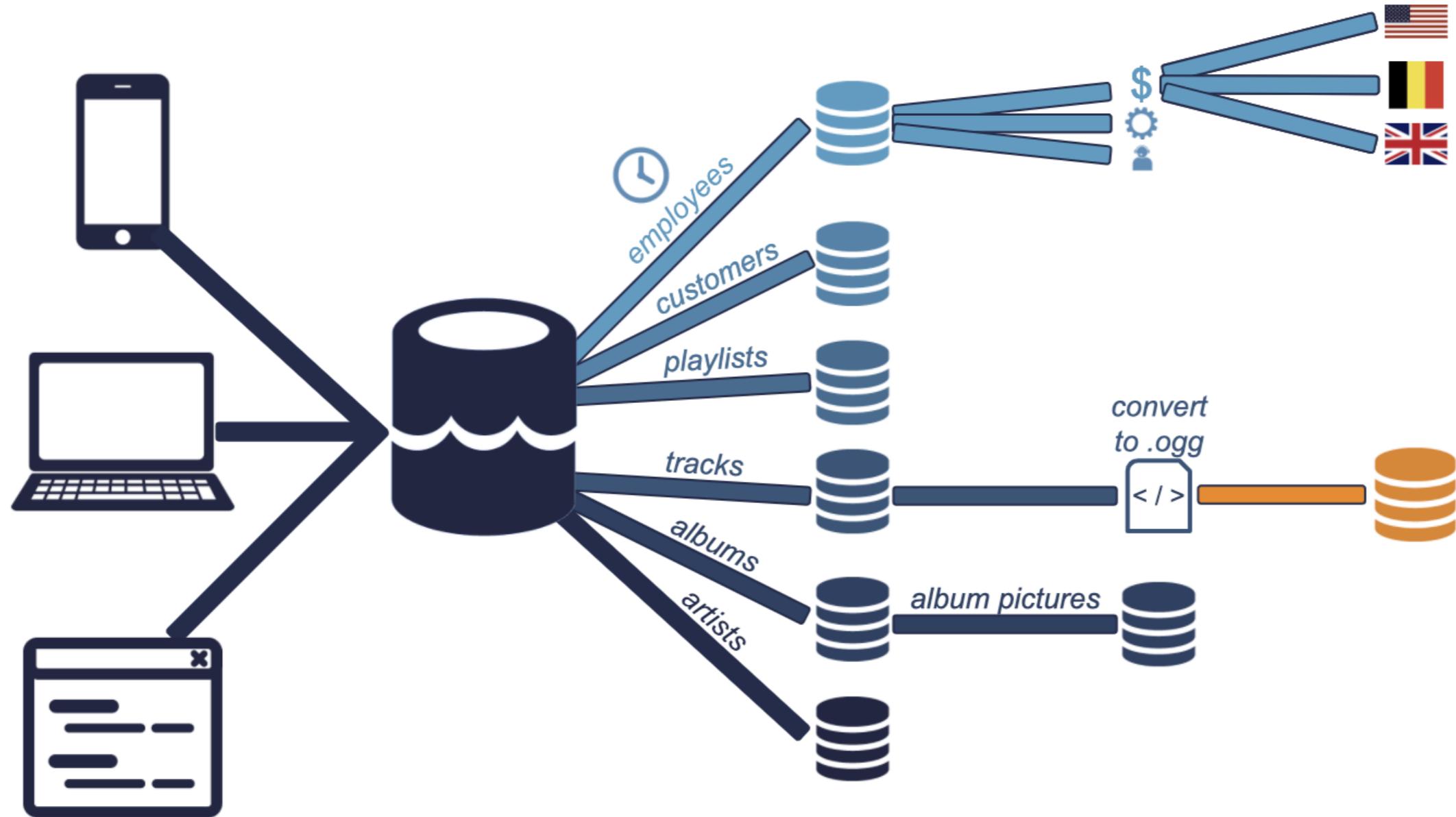
# Scheduling

- Can apply to any task listed in data processing
- Scheduling is the glue of your system
- Holds each piece and organize how they work together
- Runs tasks in a specific order and resolves all dependencies

# Manual, time and sensor scheduling

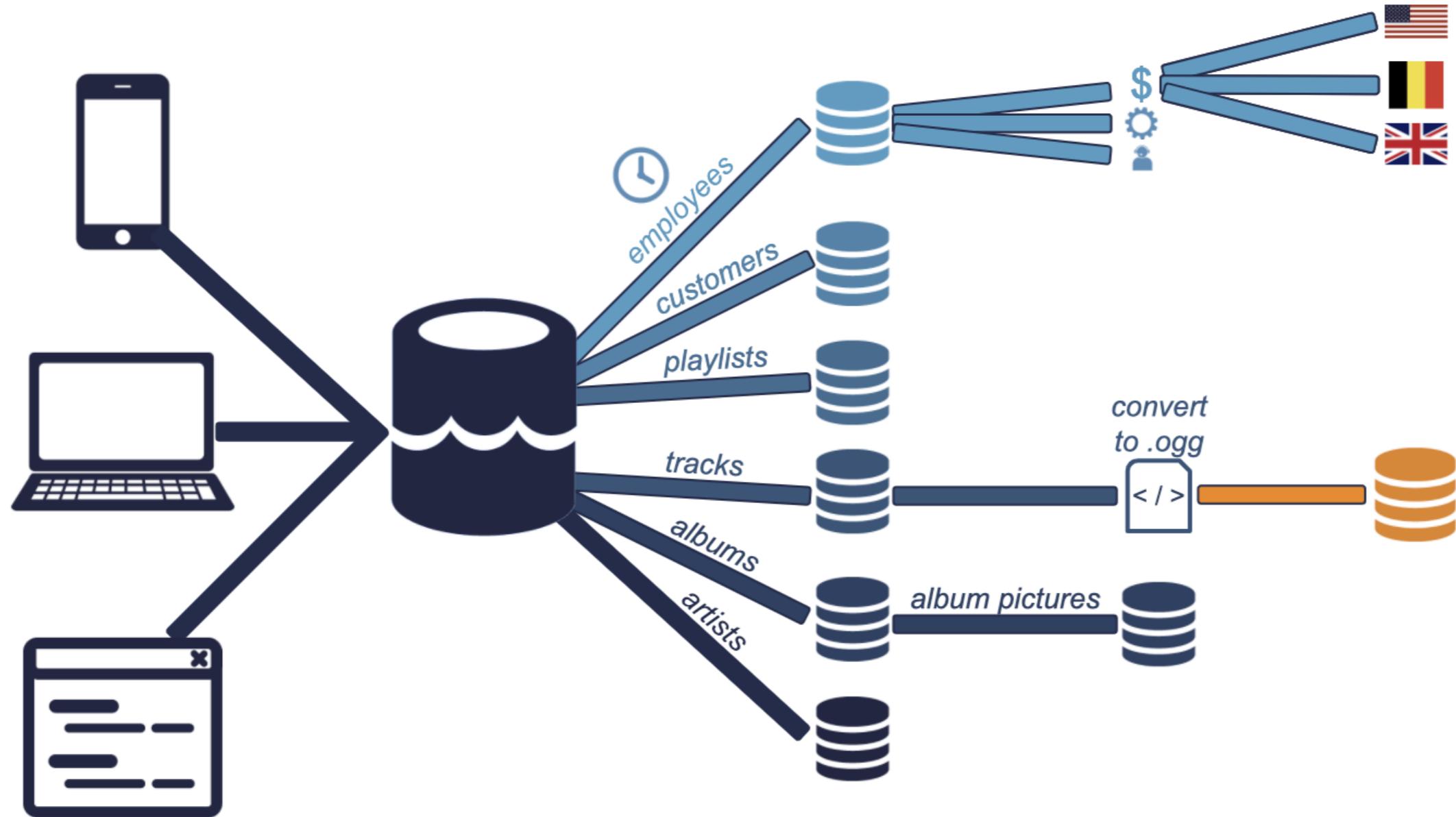
- Manually
  - Manually update the employee table

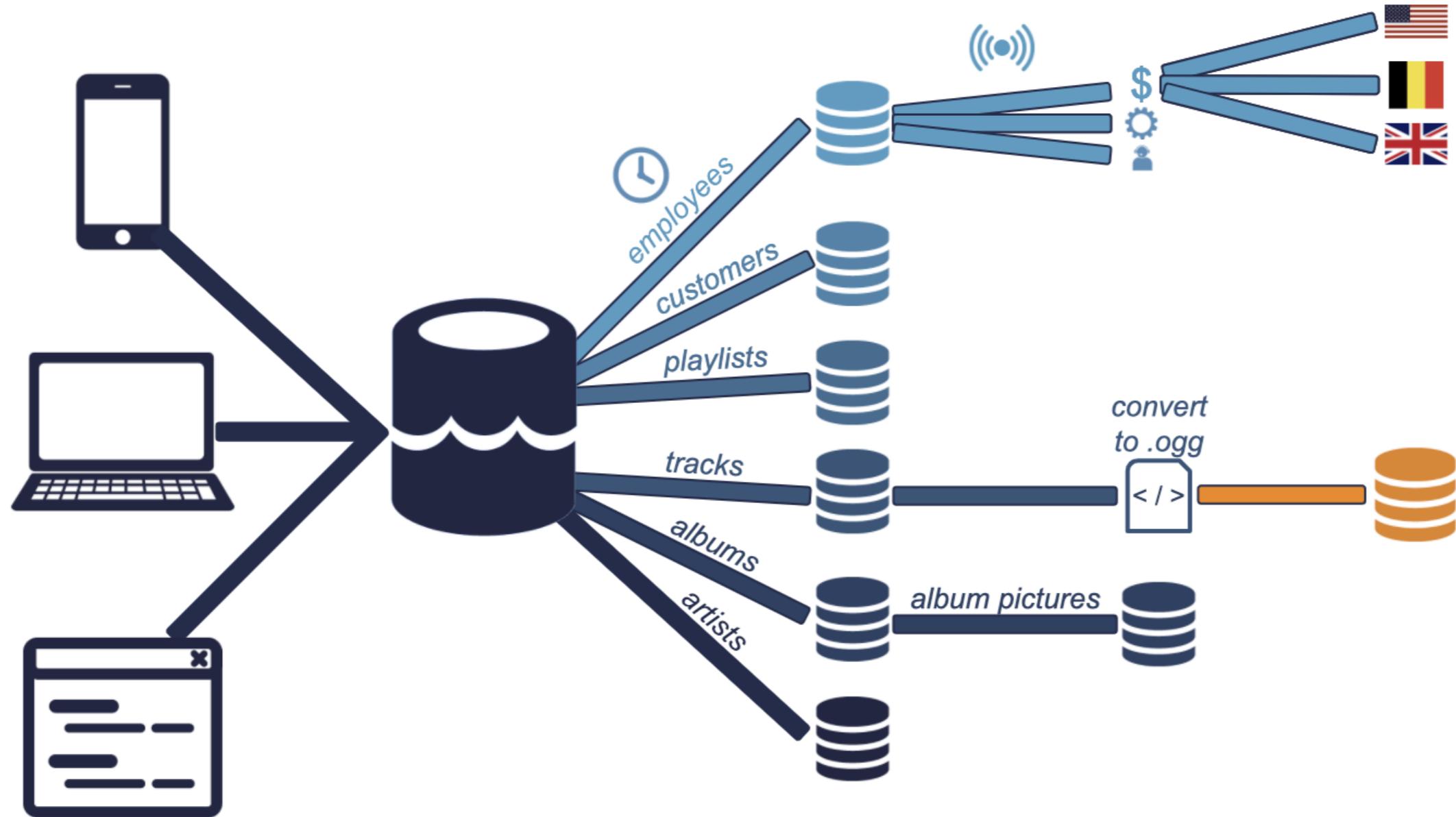




# Manual, time and sensor scheduling

- Manually
- Automatically run at a specific time
- Automatically run if a specific condition is met
  - Sensor scheduling
- Manually update the employee table
- Update the employee table at 6 AM





# Manual, time, and sensor scheduling

- Manually
- Automatically run at a specific time
- Automatically run if a specific condition is met
  - Sensor scheduling
- Manually update the employee table
- Update the employee table at 6 AM
- Update the department tables if a new employee was added

# Batches and streams

- Batches
  - Group records at intervals
  - Often cheaper
- Streams
  - Send individual records right away
  - Songs uploaded by artists
  - Employee table
  - Revenue table
  - New users signing in
  - Another example: online vs. offline listening

# Scheduling tools



# Summary

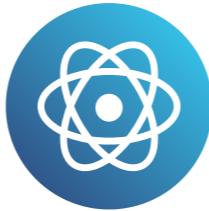
- What scheduling is
- Different ways to set it up
- Difference between batches and streams
- How scheduling is implemented at Spotflix
- Airflow, Luigi

# Let's practice!

DATA ENGINEERING FOR EVERYONE

# Parallel computing

DATA ENGINEERING FOR EVERYONE



**Hadrien Lacroix**

Content Developer at DataCamp

# Parallel computing

- Basis of modern data processing tools
- Necessary:
  - Mainly because of memory
  - Also for processing power
- How it works:
  - Split tasks up into several smaller subtasks
  - Distribute these subtasks over several computers



x 1,000

Time for  
100 t-shirts

15



x 1,000

Time for  
100 t-shirts

15



x 1,000

30



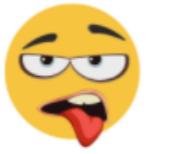
30



30



30



<sup>1</sup> Emojis by Mohamed Hassan

Time for  
100 t-shirts

15



x 1,000

30



30



30



30



Time for  
100 t-shirts

15



x 1,000

30



x 250

30



x 250

30



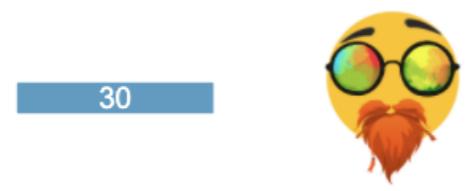
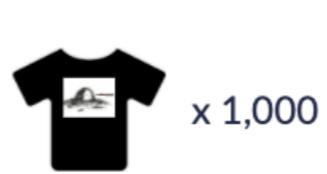
x 250

30

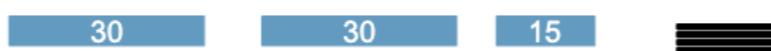
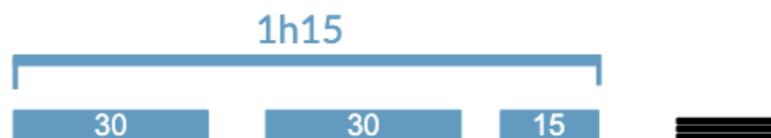


x 250

Time for  
100 t-shirts



Time for 1,000 t-shirts



Time for  
100 t-shirts

15



x 1,000

30



x 250

30



x 250

30



x 250

30



x 250

Time for 1,000 t-shirts

2h30

15 15 15 15 15 15 15 15 15



1h15

30 30 15



30 30 15



30 30 15



30 30 15



# Benefits and risks of parallel computing

- Employees = processing units
- Advantages
  - Extra processing power
  - Reduced memory footprint
- Disadvantages
  - Moving data incurs a cost
  - Communication time

Time for  
100 t-shirts

15



x 1,000

30



x 250

30



x 250

30



x 250

30



x 250

Time for 1,000 t-shirts

2h30

15 15 15 15 15 15 15 15 15



1h15

30 30 15



30 30 15



30 30 15



30 30 15



Time for  
100 t-shirts



30



30



30



30



Time for 1,000 t-shirts

2h30

15 15 15 15 15 15 15 15 15



1h15

30 30 15



30 30 15



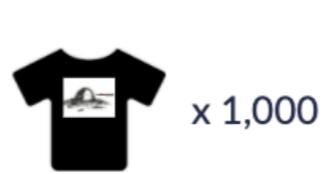
30 30 15



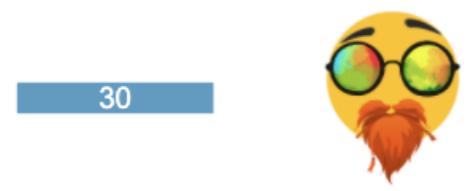
30 30 15



Time for  
100 t-shirts



x 1,000



0h10

x 250



x 250



x 250



x 250

Time for 1,000 t-shirts

2h30

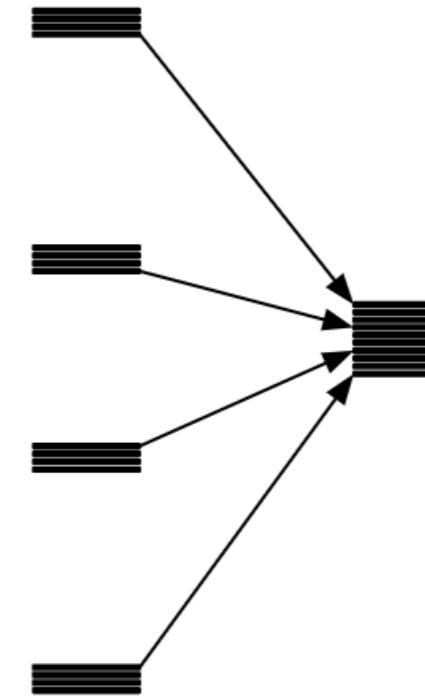
15 15 15 15 15 15 15 15 15

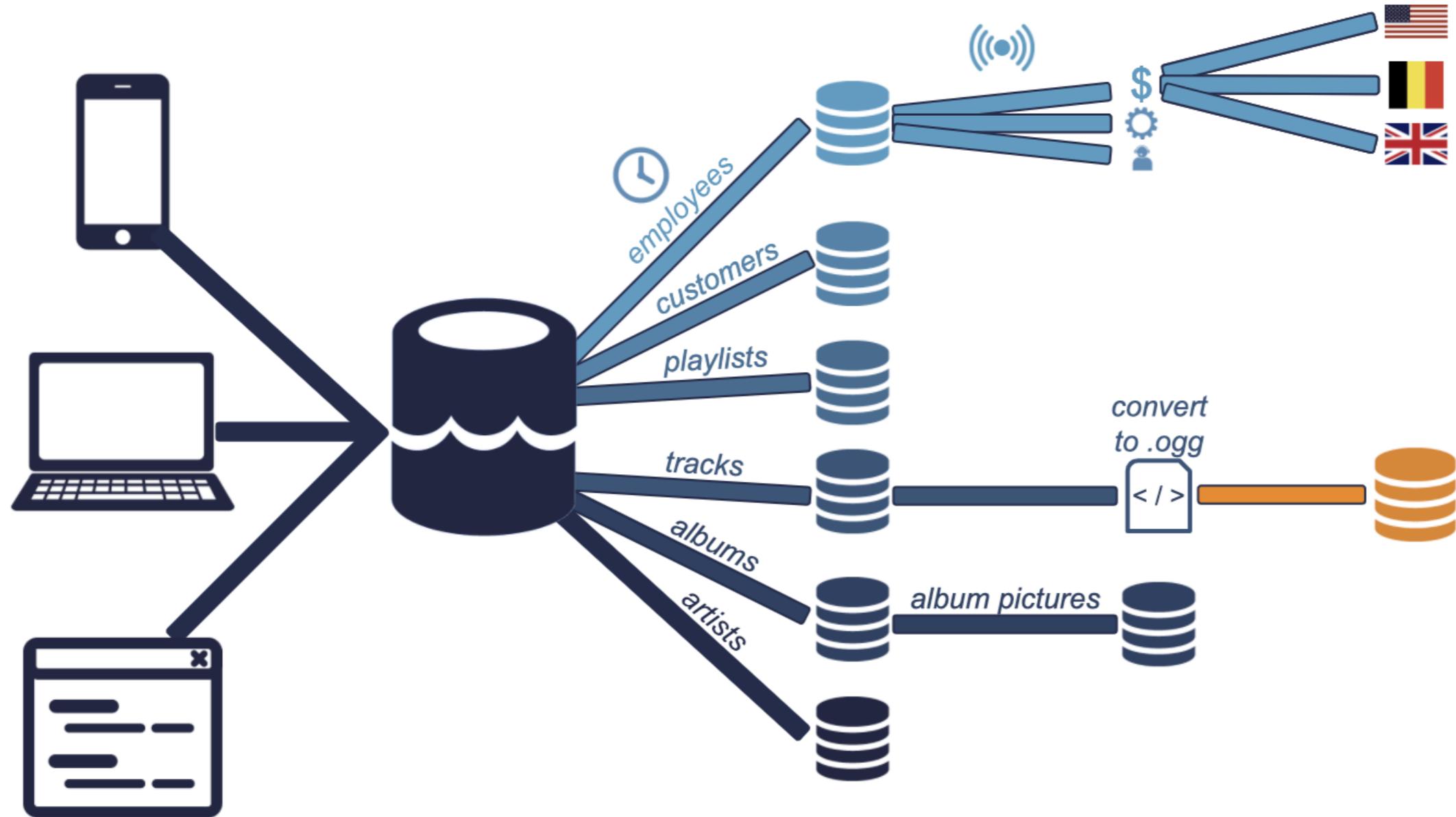


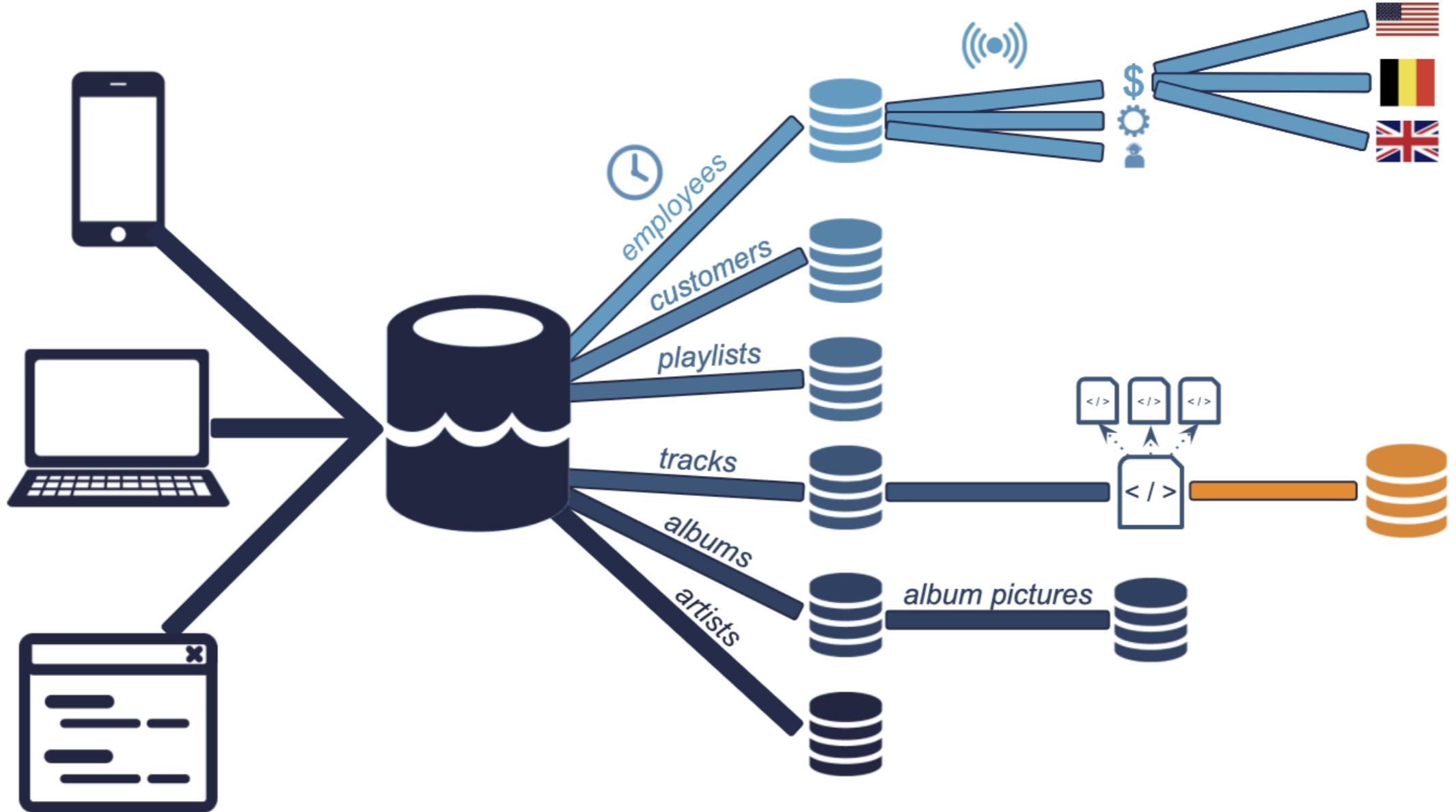
30 30 15

1h15

0h05







# Summary

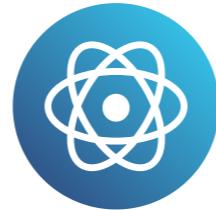
- Benefits and risks
- How it's implemented at Spotflix

# Let's practice!

DATA ENGINEERING FOR EVERYONE

# Cloud computing

DATA ENGINEERING FOR EVERYONE



**Hadrien Lacroix**  
Content Developer

# Cloud computing for data processing

## Servers on premises

- Bought
- Need space
- Electrical and maintenance cost
- Enough power for peak moments
- Processing power unused at quieter times

## Servers on the cloud

- Rented
- Don't need space
- Use just the resources we need
- When we need them
- The closer to the user the better

# Cloud computing for data storage

- Database reliability: data replication
- Risk with sensitive data



32.4%



32.4%



17.6%



32.4%



17.6%



6%

## File storage





File storage

AWS S3





## File storage

AWS S3



Azure  
Blob Storage





## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage





File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



Computation



## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2





## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2



Azure  
Virtual Machines





## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2



Azure  
Virtual Machines



Google  
Compute Engine





## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2



Azure  
Virtual Machines



Google  
Compute Engine



## Databases



## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2



Azure  
Virtual Machines



Google  
Compute Engine



## Databases

AWS RDS





## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2



Azure  
Virtual Machines



Google  
Compute Engine



## Databases

AWS RDS



Azure  
SQL Database





## File storage

AWS S3



Azure  
Blob Storage



Google  
Cloud Storage



## Computation

AWS EC2



Azure  
Virtual Machines



Google  
Compute Engine



## Databases

AWS RDS

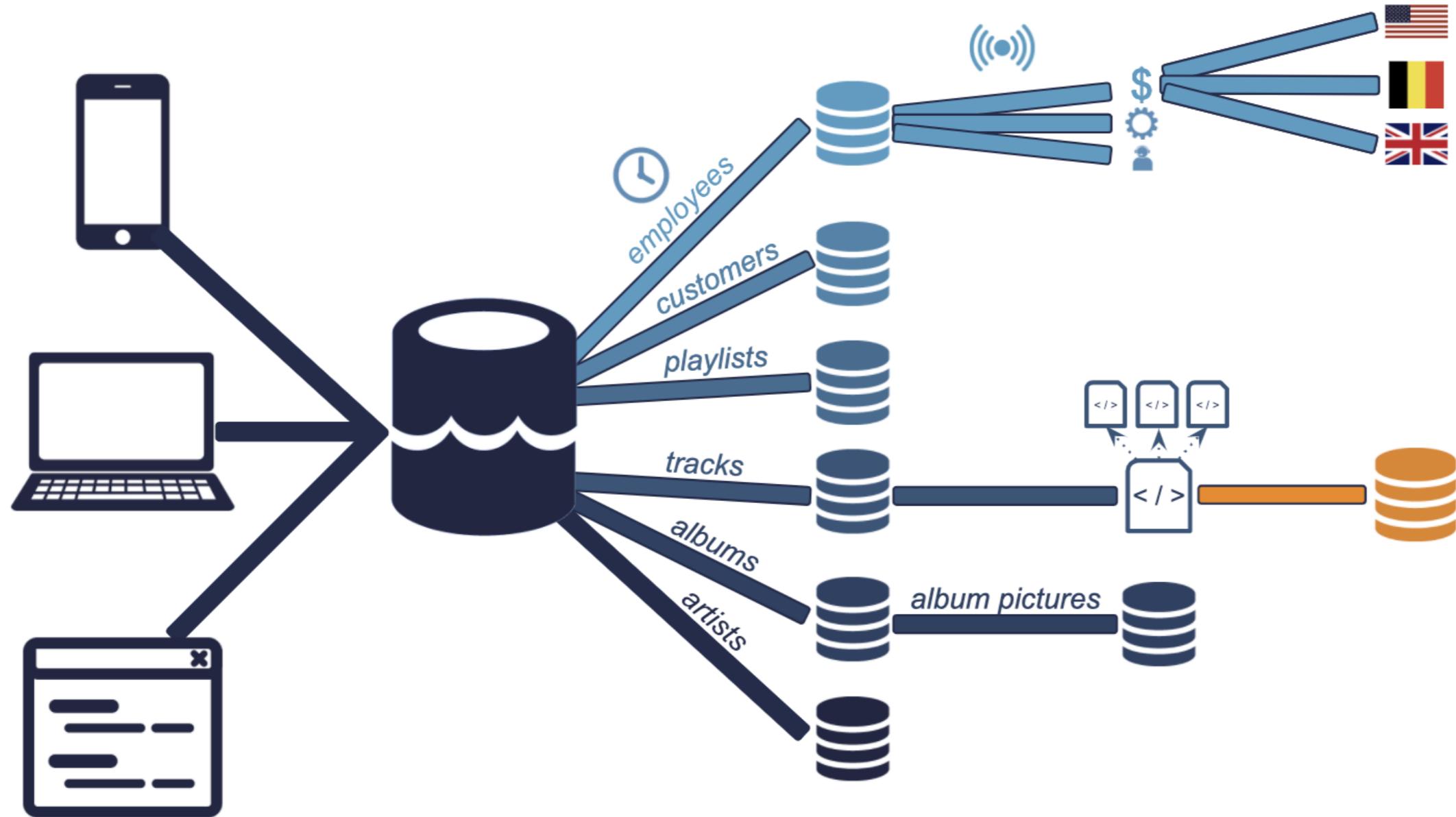


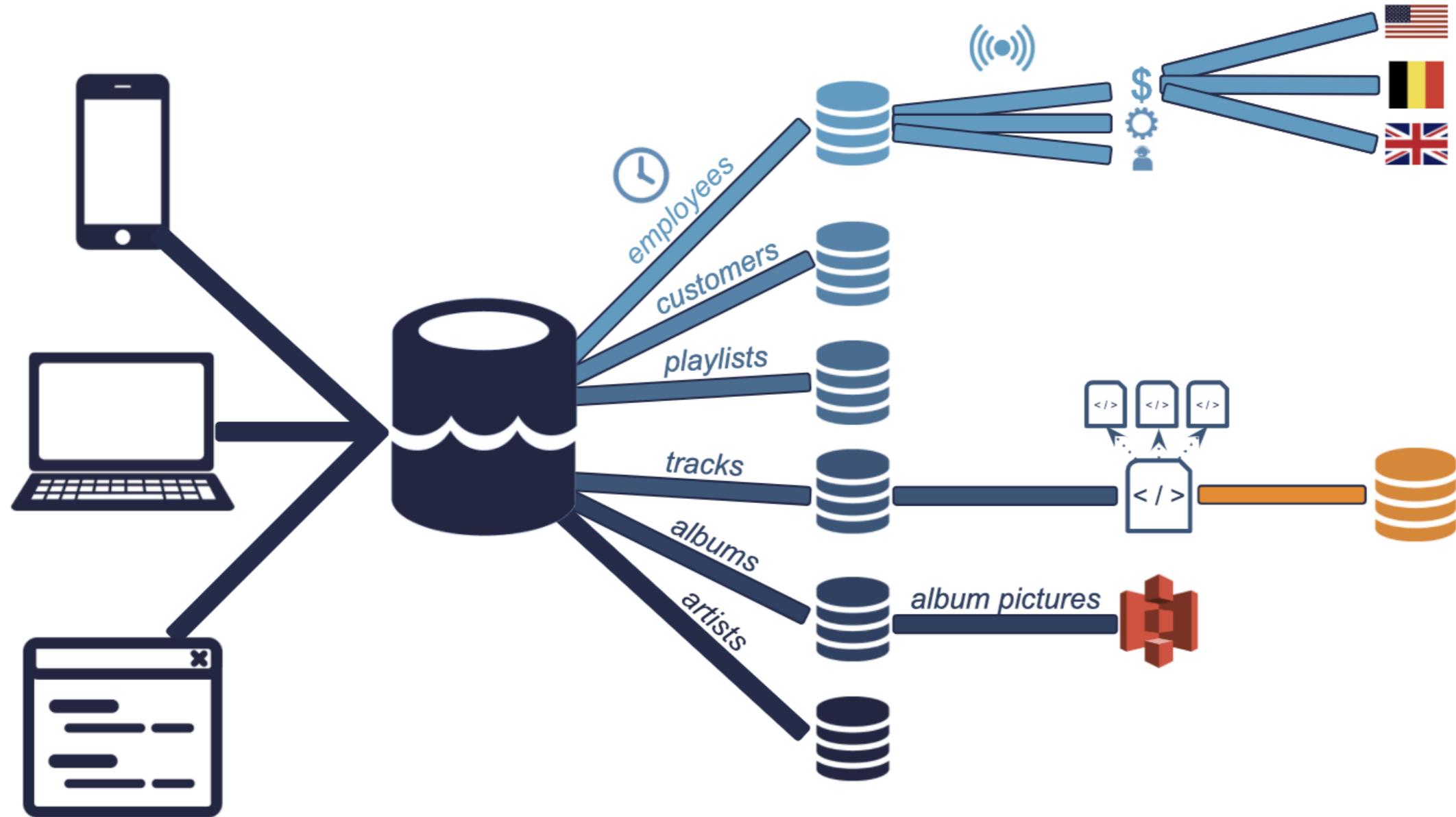
Azure  
SQL Database

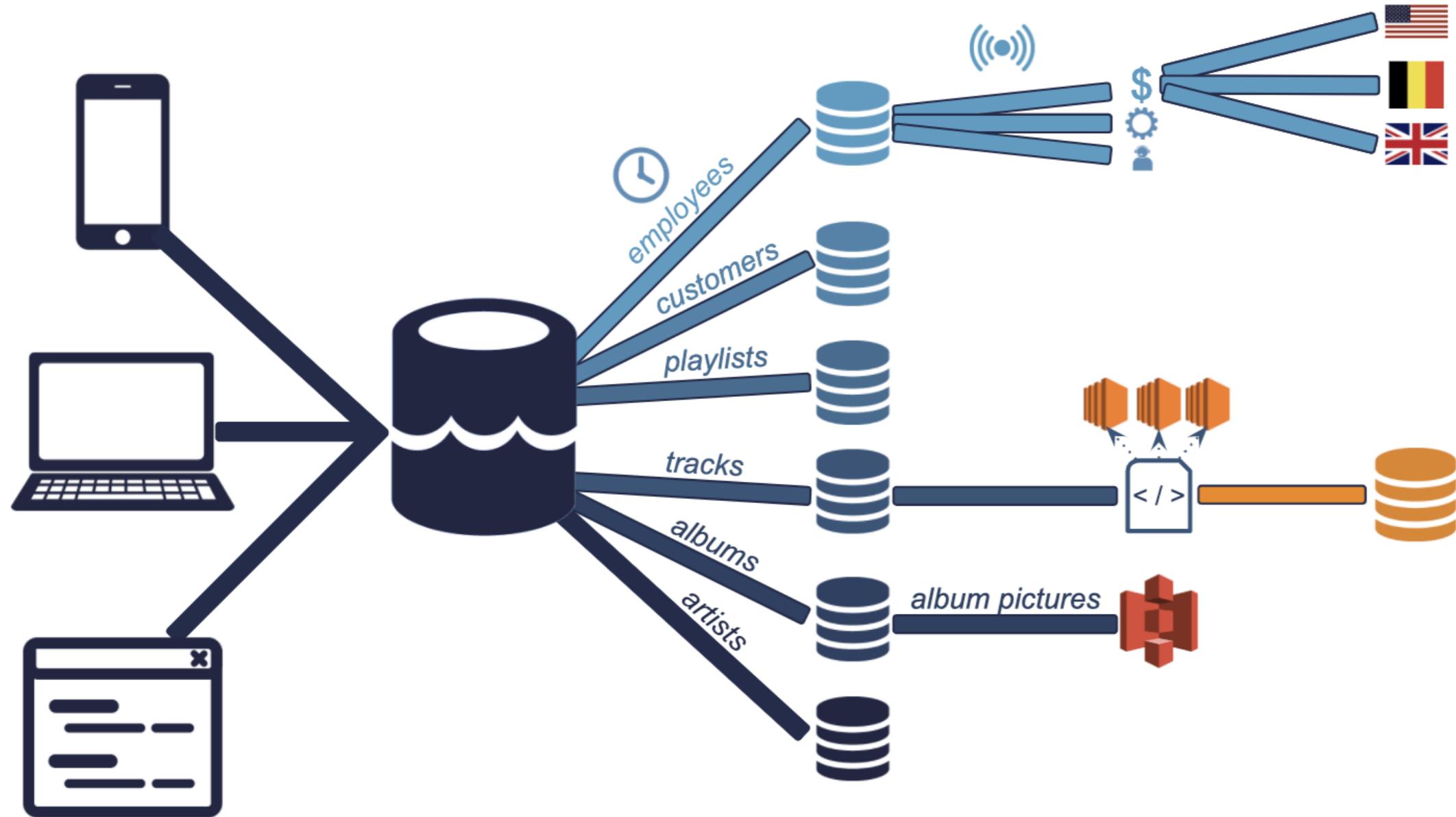


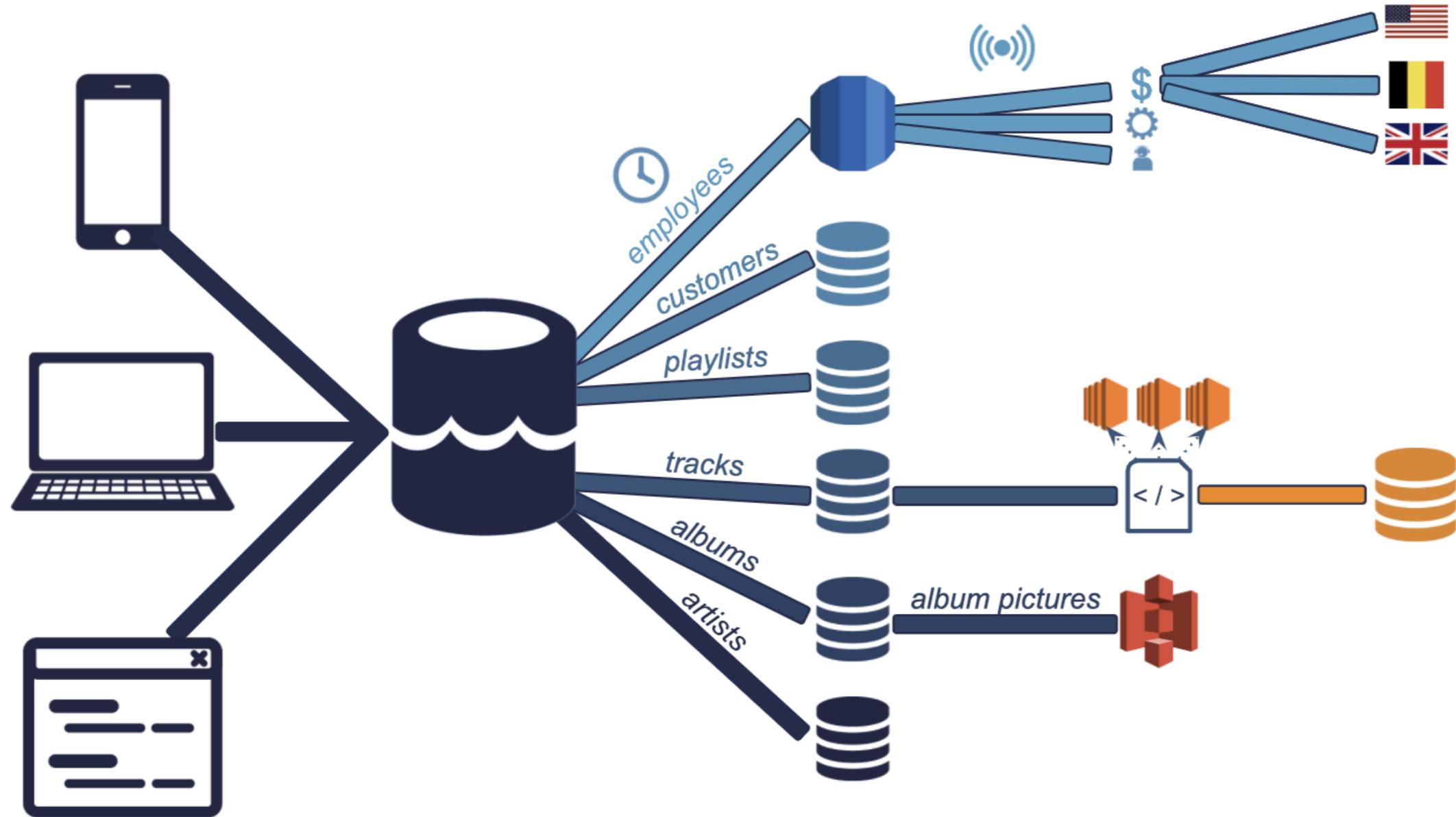
Google  
Cloud SQL











# Multicloud

## Pros

- Reducing reliance on a single vendor
- Cost-efficiencies
- Local laws requiring certain data to be physically present within the country
- Mitigating against disasters

## Cons

- Cloud providers try to lock in consumers
- Incompatibility
- Security and governance

# Summary

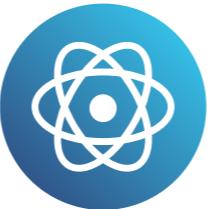
- Benefits and risks of cloud computing
- How it is implemented at Spotflix
- Can cite the main cloud providers and their services

# Let's practice!

DATA ENGINEERING FOR EVERYONE

# We are the champions

DATA ENGINEERING FOR EVERYONE



**Hadrien Lacroix**

Content Developer at DataCamp

# Actually, YOU are the champion!



# What you learned - chapter 1

- What Data Engineering is
- How important it is
- How data engineers differ from data scientists
- What a data pipeline is and how it works

# What you learned - chapter 2

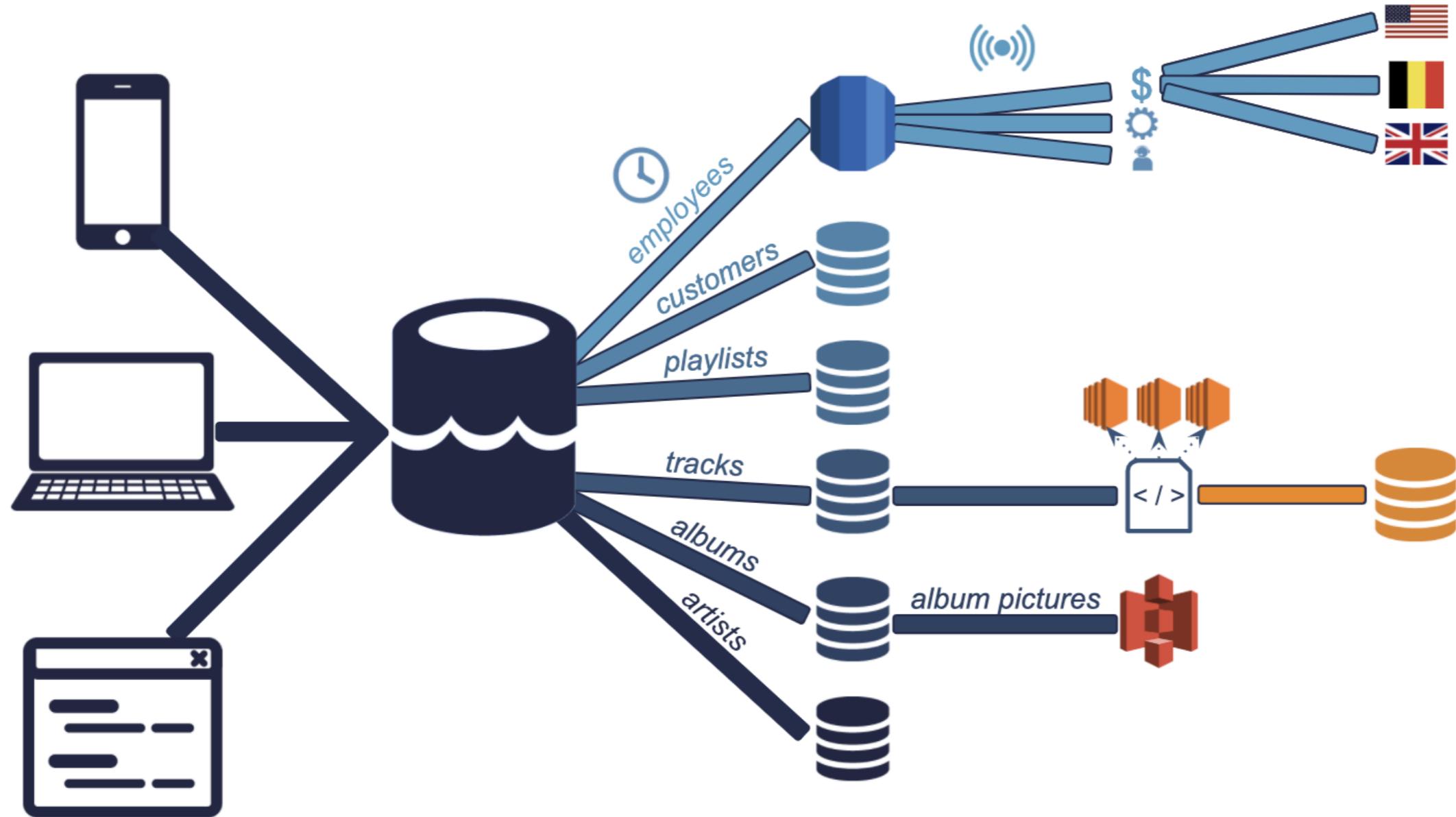
- The different structures data can take
- How fundamental SQL is
- The differences between data lakes, data warehouses and databases

# What you learned - chapter 3

- How data is processed
- How scheduling holds it all together
- Parallel computing
- Cloud computing

# And some more

- What SQL code actually looks like
- Main tools and technologies used in data engineering
- And some more



# Lexicon



My Progress

My Bookmarks

Organizations

Custom Tracks

Career Tracks

Skill Tracks

Courses

Practice

Projects

Assessments

Search Catalog

My Account



In 2019, data engineers overtook data scientists in terms of salary. How did that happen? As eager as companies are of turning their data into gold, if the mine hasn't been build, there isn't much that data scientists can surface. Data engineers lay the groundwork that makes data science possible: no wonder they are in high demand! In this conceptual course, you will get acquainted with their responsibilities, see how they differ from and enable data scientists, and how they manage the flow of data through an organization. Throughout the course, you will understand how data engineering is implemented at a fictional company named Spotflix. At the end of the course, you will be ready to have a conversation with a data engineer, understand what your company's data engineers do, or have a solid foundation to start your journey to becoming one yourself!

## 1 What is Data Engineering FREE

0%

This chapter starts with an explanation of what data engineering is and why there is an increasing need for it. Equipped with this foundational knowledge, you will then understand where Data Engineering stands in the realm of Data Science, and how a Data Engineer differs from a Data Scientist. You will finish this chapter with a first exposure to a complete data pipeline.

[VIEW CHAPTER DETAILS](#)

[Continue Chapter](#)



**Hadrien Lacroix**

Content Developer at DataCamp

Hadrien is a Content Developer at DataCamp. He's helping instructors daily build the best Data Science and Machine Learning courses possible.

[See More](#)

### COLLABORATOR(S)



**Lis Sulmont**

# A promise is a promise, DataChamps!

- All the exercises are song titles
- Search for "DataChamps" on Spotify

# Congratulations!

DATA ENGINEERING FOR EVERYONE